STATISTICAL POWER: A PRIMER

Prepared by:
The Workplace Managed Care Cross Site Evaluation Team

Christopher A. Wiesen, Ph.D.
William E. Schlenger, Ph.D.

Contact:
Jeremy W. Bray
RTI
3040 Cornwallis Road
Research Triangle Park, NC 27709-1294
(919) 541-7003
(919) 541-6683 (fax)
bray@rti.org

April 1998

Revised for the
Youth Transition into the Workplace Initiative

February 18, 2004

EXECUTIVE SUMMARY

The purpose of this paper is to provide summary information about statistical power for the collaborators in CSAP's Youth Transition into the Workplace (YIW) initiative. Power is of interest to YIW because (1) we want to have a rational basis for establishing sample sizes for the various components of the studies, and (2) we do not want to be in the position at the end of the study, if we find no difference between our "intervention" and "comparison" groups, of not being able to tell whether there truly is no difference or we just didn't have a big enough sample to detect the "true" effect.

The paper is a brief "primer" on power. It begins (Section A) with a short, nontechnical definition of power and why we should be concerned about it. Simply stated, the concept of statistical power refers to the ability of a test statistic to detect a true difference between two (or more) groups. Power is an important issue in the planning and conduct of the YIW initiative for the following reasons. First, we want to avoid concluding falsely that the YIW initiative's prevention/early intervention programs are *not* effective if in fact they *are* effective. Second, the way to reduce our chances of making such an error is to include in our studies samples that are large enough to detect differences between groups that we judge to be meaningful. Third, we do so by conducting power analyses--with each outcome included in the study--that tell us what sample sizes are required to assure a specific level of power to detect a specified effect size.

The paper then (Section B) provides a more detailed explanation of the principles underlying the concept of statistical power. These include the concept of the null hypothesis, the Type I error rate (the criterion for "statistical significance" for rejecting the null hypothesis), the "effect size" (the expected magnitude of the difference between intervention and control groups on a specified outcome variable), and the Type II error rate (probability of accepting a false null hypothesis). Using these and related concepts, the relationship between statistical power and sample size is elucidated.

Finally (Section C), the principles of power analysis are applied to the kinds of outcomes that are being studied in the YIW initiative. Tables showing sample size requirements for a variety of expected values of outcome variables.

STATISTICAL POWER: A PRIMER

The purpose of this paper is to provide summary information about statistical power and its relationship to sample size for the collaborators in CSAP's Youth Transition into the Workplace (YIW) initiative. The issue of power has arisen in a variety of contexts in discussions about design of the site-specific and the cross-site YIW studies. Power is of interest because (1) we want to have a rational basis for establishing sample sizes for the various components of the studies, and (2) we do not want to be in the position at the end of the study, if we find no difference between our "intervention" and "comparison" groups, of not being able to tell whether there truly is no difference or we just didn't have a big enough sample to detect the "true" effect.

So, we have prepared a brief "primer" on power and sample size. We begin with a short, nontechnical definition of power and why we should be concerned about it. We then provide a more thorough derivation of the concept of statistical power, for those who are interested. Finally, we address the "so what" of power by providing some estimates of sample sizes needed to detect various sizes of effect for some of the kinds of outcome variables that we will be studying.

A.   Statistical Power, Part I: The Brief Version

Simply stated, the concept of statistical power refers to the ability of a test statistic to detect a true difference between two (or more) groups. We worry about power in the YIW initiative primarily because we want to avoid getting into a situation in which we conclude, for example, that "workplace prevention programs are *not* effective in reducing substance use in the workforce (or in covered lives)," when in fact the prevention programs *did* have an impact but we did not have adequate power to detect it. Although lack of power is an important problem in any study, it is particularly embarrassing in a multisite collaborative to arrive at the end and not be able to say definitively whether or not the intervention(s) was effective.

Statistical power is influenced by several factors, including the difference between the "intervention" and "comparison" groups in a specified outcome variable (typically expressed as the "effect size"), the variance of that outcome variable, and the size of the samples. Among these, only the size of the samples is readily manipulable by the experimenter, so it is the focus of the "action" concerning power. Typically, the purpose of power analyses conducted during the design phase is to establish what size samples will be needed to assure a given level of power (minimally, 80% power) to detect a specified effect size—e.g., a pre-specified difference

between the groups, or the smallest effect that is judged to be meaningful (i.e., worth worrying about). Also, it is important to remember—particularly in the context of large, multisite research programs—that power analysis is outcome-specific (i.e., dependent on certain characteristics of specific outcomes), and therefore studies with multiple outcomes must conduct power analyses for each outcome.

So, the basic concepts are relatively simple. First, we want to avoid concluding falsely that the YIW initiative's prevention/early intervention programs are *not* effective if they in fact *are* effective. Second, one important way to reduce our chances of making such an error is to include in our studies samples that are large enough to detect differences between groups that we judge to be meaningful. Third, we do so by conducting power analyses—with each outcome included in the study—that tell us what sample sizes are required to assure a specific level of power to detect a specified effect size.

In the following section, we provide a more detailed description of statistical power and related concepts.

B.   Statistical Power, Part II: The Details

1.   The basic concepts

A statistical null hypothesis generally has the form H_0: $\theta = 0$, where θ is a parameter of interest, like a correlation coefficient, a regression parameter, or the difference between two means. We obtain a *sample estimate* of θ from observed data, for example we estimate a population mean using the sample mean. Hypothesis testing is accomplished by computing a statistic, the *test statistic*, that has a known, fixed distribution when the null hypothesis is true. This distribution is called the *null distribution*. The test statistic is usually a function of the sample estimate of θ and the estimated variance of that sample estimate.

Suppose we are comparing the means of two populations for equality, we form the null hypothesis H_0: $\mu_1 - \mu_2 = 0$ so $\theta = \mu_1 - \mu_2$. This hypothesis is generally tested by obtaining random samples from the two populations and forming an appropriate test statistic, often the t-statistic. The t-statistic is the sample estimate of $\mu_1 - \mu_2$ divided by the estimated variance of that sample estimate. If the null hypothesis is true, then this test statistic is from the null distribution, the t distribution. If the null hypothesis is false, the test statistic is a random variable from a different distribution, referred to as the *alternative distribution*. The alternative distribution is usually not known, but may be estimable, and it always differs from the null distribution. By comparing the value of the test statistic to the null distribution, we can state the

probability of observing that statistic, or one more extreme (further from 0), if the null hypothesis is true. If that probability is very low, we reject the null hypothesis in favor of the *alternative hypothesis*, that is, H_a: $\theta \ne 0$.

Unfortunately, whether the null hypothesis is true or false, the test statistic can take on the same range of values (although with different probabilities), so we may never say with certainty that the null hypothesis is false, no matter how extreme the statistic is. So while it is very unlikely that we would observe a very extreme value for the test statistic if the null hypothesis is true, it is not impossible. We must allow some probability of rejecting the null hypothesis when it is true, in order to allow the detection of true non-null effects. The probability of rejecting a true null hypothesis is called the Type I error rate or α, and is generally set to some arbitrarily small value like .05.

Associated with the null distribution is the *critical value*. Values more extreme than the critical value fall in the *rejection region*, and have a probability less than α of being observed in the null distribution. If the test statistic is in the rejection region, then we reject the null hypothesis; if not, we do not reject the null hypothesis. The probability of failing to reject a false null hypothesis the Type II error rate or β. 1-β, the probability of rejecting a false null hypothesis, is called power. Power is a direct function of the degree to which the null and alternative distributions overlap (less overlap = more power) and α.

Since the null distribution is fixed and α is generally set by convention to .05, only the alternative distribution may by influenced by the experimenter. Two factors influence the alternative distribution. One is the *magnitude of effect*. Magnitude of effect is the degree to which the null hypothesis is incorrect, or how far the true value of θ is from 0. Larger magnitudes of effect lead to greater disparity between the null and alternative distributions and more power. The other is the variance of the sample estimate of θ. Sample estimate variation may be thought of as follows. Suppose we want to estimate μ, the mean of a population. We would accomplish that by obtaining a sample of size n and computing the mean of that sample. Suppose that we were to obtain a large number of samples, all of size n, and use each one to estimate μ. These values would all be equally good estimates of μ, but they would not all be equal to each other. The degree to which they vary is the variance of the sample estimate. A researcher will usually only obtain one sample, but a single sample is usually sufficient to get a good estimate of the variance of the estimate of θ. Smaller sample estimate variation results in greater power.

In observational studies, the magnitude of effect is usually beyond the control of the researcher. This leaves the variance of the sample estimate of θ as the only factor which affects power that the researcher may control. The variance of the sample estimate is generally a function of the *population variance* (roughly, how much population values vary around their mean) and the sample size. If there is high variation around the mean within a population, then we would observe large variation in estimates across samples. Conversely, small population variance will result in sample estimates that have lower variation. Independent of population variation, bigger samples result in lower sample estimate variance. With larger samples, observations far from the mean in one direction tend to be canceled out by observations extreme in the other direction so estimates become more precise.

2.   *Estimating power and required sample size*

Experimenters are often asked to estimate the power for a particular experiment. In order to accomplish that, the experimenter must have four pieces of information. The first and most important is a well defined null hypothesis. Power is only defined in the presence of a meaningful null hypothesis and without that, good power estimates are impossible. Second is an estimate of the magnitude of effect. This may be an actual value such as $\mu_1 - \mu_2 = 4$ or more general, such as "large" or "medium" effects. Third, is the population variance and forth is the sample size.

A well defined hypothesis and the sample size are fixed by the researcher. The magnitude of effect and population variance are usually not exactly known nor are they always manipulable. Sometimes pilot testing or previous research yields reasonable estimates of both magnitude of effect and population variance. These estimates may be treated as true population values in estimating power. The experimenter may define a *meaningful effect*, that is a point below which the discrepancy between the null and alternative hypothesis, although real, would not be of substantive interest. For example, the researcher may say that if the difference between two population means is less than 4, then the effect is trivial and essentially the same as if two means were equal. The researcher would then set the magnitude of effect to 4. Another alternative is to define the magnitude of effect in terms of the unknown population variance, or its square root, the standard deviation. The experimenter may define a "medium" effect as, say, $\mu_1 - \mu_2 = 1.5$ standard deviations and a "large" effect as $\mu_1 - \mu_2 = 2$ standard deviations. Effects defined this way satisfy both the magnitude of effect and variance requirements.

Given a well defined null hypothesis, and good estimates of the magnitude of effect and variance, a researcher may want to estimate power for a range of possible sample sizes. Alternatively, the researcher may want to estimate the sample size necessary to obtain a specified level of power. In either case, there are equations for estimating one of the variables, if the others are given. Figure 1 is a graphic depiction of a comparison between hypothetical null and alternative distributions. The left-hand curve is the null distribution, the distribution that the test statistic will have if the null hypothesis is true. For any statistical test, the null distribution is fixed and its properties are well understood. The other curve is the alternative distribution, the distribution of the test statistic for a certain situation when the null hypothesis is false. The alternative distribution shown is just one of an infinite number of possible null distribution under the infinite number of possible non-null situations. Other alternative distributions will have more or less overlap with the null distributions. The x-axis is the value of the test statistic. At the center of the null distribution, the value of the test statistic is zero.

The critical value is chosen to divide the null distribution into two regions, the portion of the null distribution to the right of the critical value is α, usually 5%. If the null hypothesis is true, then the probability that the test statistic is more extreme then the critical value is α. The values along the x-axis more extreme then the critical value constitute the rejection region. The critical value also divides the alternative distribution. The portion to the left is β, the probability of accepting a false null hypothesis. The portion to the right is power, the probability of properly rejecting a false null hypothesis. Increased sample sizes increase the disparity between the null and alternative hypotheses, increasing power.

C.   Statistical Power, Part III: Sample Size Estimates

So what does this mean for YIW collaborators? To answer this question, we begin by looking at the kinds of outcomes (dependent variables) that we will be studying, both in the cross-site and in the individual studies.

Based on our review of the sites' applications, the discussions we've had with each site about interventions and design, and the logic models that the sites have developed, it is clear that one major class of dependent variables are prevalence rates or other proportions—e.g., the prevalence of "binge drinking" in the workforce (or among covered lives), the proportion of employees (or covered lives) who had an emergency room visit during a specific time period. Table 1 shows for varying proportions/percentages the sample sizes required for 80% power (one-tailed test) to detect relatively small, medium, and larger effects.

The sample sizes shown in Table 1 are relatively large, particularly for the smaller proportions and smaller effect sizes. For example, if the prevalence of "binge drinking" were 5%, and we wanted to have 80% power to detect a 1-percentage point reduction in that prevalence in the intervention vs comparison group (i.e., 5% vs 4%), we would need a sample of 6,391 participants in each group! This demonstrates an unfortunate reality of statistical power for endpoints that are expressed as proportions—the smaller the prevalence, the more subjects required. By way of comparison, suppose that the prevalence of "negative attitudes toward substance abuse" was 40%, and we wanted to have 80% power to detect a 20-percentage point difference between intervention and comparison groups (i.e., 40% vs 20%). In this case, we would need only 76 subjects per group.

These analyses underscore the need to think through the research and policy questions that underlie the YIW initiative, and to plan carefully that we will be able to address those questions definitively. The primary point is: given the kinds of interventions that are being tested in the YIW initiative, and therefore the kinds of effect sizes we are likely to have, are the current estimates of sample size realistic? Over the coming weeks, the Cross-Site Evaluation Contractor will be talking with each of the site teams about a variety of issues, including sample size, to examine the implications for each of the sites' studies. As we have emphasized in our discussions of cross-site issues at past Workplace Managed Care Steering Committee meetings, there are a number of factors other than sample size that influence Type II error (e.g., design-based influences, measurement reliability, analytic method). Sample size, however, remains one of the most important determinants.

Table 1. **Sample sizes required for varying outcome proportions and expected differences between intervention and comparison groups.**

If the *Outcome* is:	And we want to detect a *Difference* of:	Sample Size *Required**
5%	1%	6,391
	2%	1,725
	3%	821
10%	1.5%	5,236
	3%	1,388
	5%	534
15%	3%	1,882
	6%	502
	9%	236
20%	2%	5,109
	5%	857
	10%	229
25%	25%	3,818
	5%	980
	12.5%	168
30%	3%	2,951
	9%	342
	15%	127
35%	3.5%	2,337
	9%	362
	17.5%	98
40%	4%	1,873
	10%	304
	20%	76
45%	4.5%	1,514
	10%	307
	22.5%	59

*Sample size <u>per group</u> required to achieve a power of .80 in a one tailed test.